Neutral components show a hierarchical community structure in the genotype-phenotype map of RNA secondary structure

Marcel Weiß[1,2] and Sebastian E. Ahnert[3,4]
[1] *Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, Cambridge, UK*
[2] *Sainsbury Laboratory, University of Cambridge, Cambridge, UK*
[3] *Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK*
[4] *The Alan Turing Institute, British Library, London, UK*
(Dated: September 29, 2020)

Genotype-phenotype (GP) maps describe the relationship between biological sequences and structural or functional outcomes. They can be represented as networks in which genotypes are the nodes, and one-point mutations between them are the edges. The genotypes that map to the same phenotype form subnetworks consisting of one or multiple disjoint connected components – so-called neutral components (NCs). For the GP map of RNA secondary structure, the NCs have been found to exhibit distinctive network features that can affect the dynamical processes taking place on them. Here, we focus on the community structure of RNA secondary structure NCs. Building on previous findings, we introduce a method to reveal the hierarchical community structure solely from the sequence constraints and composition of the genotypes that form a given NC. Thereby, we obtain modularity values similar to common community detection algorithms, which are much more complex. From this knowledge, we endorse a sampling method that allows a fast exploration of the different communities of a given NC. Further, we introduce a way to estimate the community structure from genotype samples, which is useful when an exhaustive analysis of the NC is not feasible, as it is the case for longer sequence lengths.

I. INTRODUCTION

The mapping between genotypes and phenotypes can be understood as a network, and can therefore be analysed using the tools of network science. The fundamental idea [1–3] is to consider the space of all genotype sequences – for example all RNA, DNA, or amino acid sequences of a given fixed length – as a network, in which each sequence is represented by a node and each one-point mutation between two genotypes as an edge. Genotypes that map to the same phenotype have been shown to be correlated [4] and to form subnetworks. These subnetworks – commonly referred to as neutral sets or neutral networks – are either fully connected or consist of multiple disjoint components – commonly referred to as neutral components (NCs). One-point mutations between genotypes in a NC do not change the phenotype (and therefore the fitness) and are labelled as neutral. By contrast, mutations between genotypes of disjoint NCs of the same phenotype can involve genotypes with phenotypes of lower fitness, meaning unfavourable steps from an evolutionary perspective. Thus, in this article, we will focus on NCs – the most essential neutral units for evolving populations on the genotype space.

One of the most extensively studied genotype-phenotype (GP) maps is the mapping between RNA sequences (genotypes) and their secondary structure (phenotypes) [4–13]. The secondary structure is an abstraction of the full three-dimensional spatial structure and only considers the base pair configuration – often shown using the 'dot-bracket notation'. If we compare the set of genotypes that belong to a given NC, the unpaired sequence sites tend to be the most unconstrained, meaning that they allow a significant number of neutral mutations so that a range of different letters (nucleotides) can be found at each of these sites. By contrast, the paired sites are mostly constrained and only allow a limited number of neutral mutations so that only a limited range of different letters is found at each of these sites.

Aguirre et al. [9] were the first to thoroughly analyse the topological properties of the NCs of the RNA secondary structure GP map of sequence length $L = 12$. Among other unique properties, they showed that the NC networks are assortative and exhibit a community structure. For an example NC, they studied the community structure by comparing the degree of the nodes to their eigenvector centrality. In this context, they discovered that these communities can be characterised by the letter combinations at the base pairs in the stacks of the respective phenotype [9]. *Capitán et al.* [14] continued this research and further studied the community structure of RNA secondary structure NCs. Using another example NC, they identified 'dynamical communities' and a hierarchical community structure by running population dynamics on the NC. In particular they studied the equilibrium distribution – related to the eigenvectors of the adjacency matrix also considered in [9] – as well as the time to equilibrium distribution. Their findings confirmed that the identified communities can be characterised by letter combinations at certain sequence sites, in this case at two unpaired sites, since the paired sites are fully constrained for the considered example NC. A further study of the topological properties of NCs can be found in the PhD thesis by *Greenbury* [15]. Studying the GP maps of RNA secondary structure and of the so-called Polyomino model [12, 16], this work also draws links between the community structure of a NC and the sequence constraints and composition of the genotype sequences that form this NC. Building on empirical data, in [17],

Aguilar-Rodríguez et al. study the architecture – including the community structure – of genotype networks in the GP map of transcription factor binding sites. Studies on the impact of topological properties of NCs on population dynamics range from initial work by *van Nimwegen et al.* [18] to more recent studies by *Manrubia* and co-workers [19, 20].

In this article, we present a method that reveals the hierarchical community structure of a NC solely from the sequence constraints and composition of its genotypes, without the necessity of running complex community detection algorithms. In the first instance, this approach is designed to reveal the community structure of NCs that are fully mapped, meaning that we know all its constituent sequences. However, we also introduce a way to estimate the community structure from genotype samples of the full NC. Here, we only consider the RNA secondary structure GP map, but our framework could potentially be adapted to many other GP maps.

As the main model system, we use the RNA secondary structure GP map of sequence length $L = 12$, for which the NCs have a size that allows meaningful plots of the NC networks as well as a network analysis in a reasonable computational time. To predict the secondary structure of a given RNA sequence, we use the Python implementation of the so-called *ViennaRNA* package [21–23] (version 2.4.9, default parameters) with its function *RNA.fold*. There are 431 NCs (ignoring the undefined phenotype, or unbound structure), which we rank with respect to their size, with the largest receiving rank 1. For the network analysis, we mainly use the Python package *NetworkX* (version 2.1).

The article is structured as follows. We start by explaining our method to reveal the hierarchical community structure of a NC and apply it to the NCs of the $L = 12$ RNA secondary structure GP map. This is followed by the endorsement of a sampling method that allows a fast exploration of the different NC communities. Finally, using this method, we introduce a way to estimate the community structure from samples of genotypes. We apply it to two NCs comprised of naturally occurring functional non-coding RNA sequences of longer lengths.

II. SEQUENCE-BASED COMMUNITIES

We begin by introducing our method to reveal the hierarchical communities of the NCs that can be applied whenever the NC of interest is fully known.

A. Method

The starting point is the number of neutral mutations per sequence site averaged over the NC – a measure of the average constraint of sequence sites across the genotypes of the NC. For an individual genotype, the number of neutral mutations for a particular site measures how many different letters the current letter at this site can be mutated to without changing the phenotype. For RNA with its four letter alphabet, it ranges from 0 (fully constrained) to 3 (fully unconstrained).

In order to reveal the communities, we consider all sites, starting with the most constrained (smallest average number of neutral mutations) and moving to the least constrained (largest average number of neutral mutations). We exclude fully constrained sites (zero average number of neutral mutations) as all genotypes of the NC have the same letters at these sites. We begin with an empty list of constrained positions and add to this list the position of the most constrained site with non-zero average number of neutral mutations. Next, we go through the genotypes of the NC and associate them to communities according to the letters that they have at these positions. In other words, the different communities are defined by different letter combinations at the constrained positions and genotypes belong to the same community if they have the same letters at these positions. Once all genotypes of the NC have been considered, we add the position of the next most constrained site to our list of constrained positions and repeat the procedure and so on until we have added the positions of all sites. Whenever multiple sites have exactly the same average number of neutral mutations, we add their positions in the same step. Figure 1 illustrates the procedure.

This association of communities with letter combinations at constrained positions is similar to how *Manrubia* and co-workers characterise the communities for two individual RNA secondary structure NCs in [9] (letter combinations at paired sites) and [14] (letter combinations at constrained unpaired sites while the paired sites are fully constrained). Our method unifies these ideas, as well as the idea by *Greenbury* [15] to relate sequence constraints and composition to the community division of a NC, by providing a simple algorithm that can be applied to any fully known NC.

For every step in our method, we calculate the modularity of the discovered community structure. The concept of modularity in this context was introduced by *Newman* and *Girvan* [24, 25] and measures whether the sets of nodes in a given partition are more densely connected inside each set than one would expect by chance. The modularity Q is calculated by [24, 25]:

$$Q = \sum_i \left(e_{ii} - a_i^2\right) \tag{1}$$

where the sum runs over all communities. e_{ii} measures the fraction of all ends of edges that are attached with both ends to nodes in community i and a_i measures the fraction of all ends of edges that are attached to nodes in community i [24, 25]. The contribution of a given community i to the modularity is therefore positive if e_{ii} is greater than the null expectation a_i^2. The higher the modularity (which ranges from -1 to 1), the more meaningful a community division.

With each step in our procedure, a finer-grained community structure is revealed. At a fully constrained site,

there can only be one letter. For RNA, if a site with non-zero average number of neutral mutations is unpaired, in principle, there can be four different letters, if it is paired, two different letters, and for two sites corresponding to a same base pair, three different letter combinations. This sets an upper limit for the number of communities that are possible in principle. Therefore, it is to be expected that the modularity initially will increase when adding more sites to the list of constrained positions, will reach a maximum that corresponds to the most meaningful community structure, and then will decrease as the community structures become too fine-grained.

The association of communities with letter combinations at constrained positions also allows a coarse-grained network representation of the full NC topology. We simply associate a given letter combination with a coarse-grained node and connect two coarse-grained nodes whenever the associated letter combinations differ by only one letter. We choose the relative size of a coarse-grained node according to the number of genotypes within the respective community. There are a few caveats: Firstly, nodes in the original NC network with the same letter combinations at the constrained positions and therefore within the same community might not be fully connected, which is not reflected by representing all of them by one coarse-grained node. Secondly, there might not be any connection between nodes in the original NC network corresponding to two different coarse-grained nodes, even if the associated letter combinations at the constrained positions differ by only one letter. If necessary, this issue could be resolved by referring to the edges in the original NC network.

B. Examples

In the following, with Figure 1, we explain the method using four example NCs from the $L = 12$ GP map that highlight different cases in a particularly clear way. For all examples, we show the average number of neutral mutations per site as well as for several steps of the method, the NC network with coloured communities, its modularity, its coarse-grained representation, and in some cases the associated letter combinations at the constrained positions. Both the full and the coarse-grained networks are plotted using a force-directed graph layout algorithm.

The first example in Figure 1 (A) is the NC of rank 32 with a phenotype that has three base pairs. The force-directed layout shows a clear community structure and suggests the presence of seven distinct communities. The two sites corresponding to the outermost base pair are the most constrained. In addition, they have the same constraint. Therefore, both positions are considered in the first step. At these two positions, we find three different letter combinations: The two Watson-Crick base pairs CG and UA and a wobble base pair UG connecting both. This means that we reveal three communities at this step. In the second step, we add the positions of the two sites corresponding to the innermost base pair, which again have the same constraint. At these four positions, we find five letter combinations, i.e. five communities. After the fourth step, the positions of all paired sites are added and we reveal seven communities exactly matching those that the force-directed layout suggests. The central community corresponds to three CG or GC base pairs, the most stable base pairs. The three communities adjacent to it correspond to letter combinations where one of those base pairs is exchanged for a UG or GU wobble base pair. Adjacent to each of these three communities, respectively, there is a community corresponding to a letter combination where the wobble base pair is exchanged for a UA or AU base pair. In the fifth step, the position of an unpaired site is added, which corresponds to a significant decrease in the constraint (increase in the average number of neutral mutations). We find that each of the communities from the previous step splits into three or four subcommunities – a new hierarchy layer. We stop at this step as further steps would simply lead to even more fine-grained and less meaningful community structures. This example shows how hierarchy layers are related to the site constraints. Sites that have a roughly similar constraint create a particular layer in a hierarchy. Adding the position of one of these sites progressively builds up such a layer, as we see here until step four. A significant decrease in the constraint of an added site reflects a change of the hierarchical layer, as we see here for step five. As already suggested by the force-directed layout, the maximum modularity is reached for step four, though it only slightly decreases for step three and five, but then further beyond that.

The second example in Figure 1 (B) is the NC of rank 36, which corresponds to a two base pair phenotype. All paired sites are fully constrained and all genotypes have the same letters at these positions. In the first two steps, we add the positions of two unpaired sites that are roughly similarly constrained. We find ten different letter combinations, and the communities match with the force-directed layout. The communities are less pronounced in the layout, probably because the unpaired sites considered here are less constrained than the paired sites considered in the previous example. In the third step, the position of a significantly less constrained site is added. This leads to an observable change of the hierarchy layer as explained. The maximum modularity is found for step two and slightly decreases for step three, but does so more significantly for further steps. This NC and the community structure for step two resemble the example discussed in [14].

The third example in Figure 1 (C) nicely displays two observable changes of the hierarchy layer. It is the NC of rank 41 with a phenotype that has three base pairs. In contrast to the first example, two of the unpaired sites are significantly more constrained than the other ones, meaning three clusters of sites: Six roughly similarly constrained paired sites, two roughly similarly half-constrained unpaired sites and four roughly similarly unconstrained unpaired sites. As explained, an observable change of the hierarchy layer occurs for step five and seven,

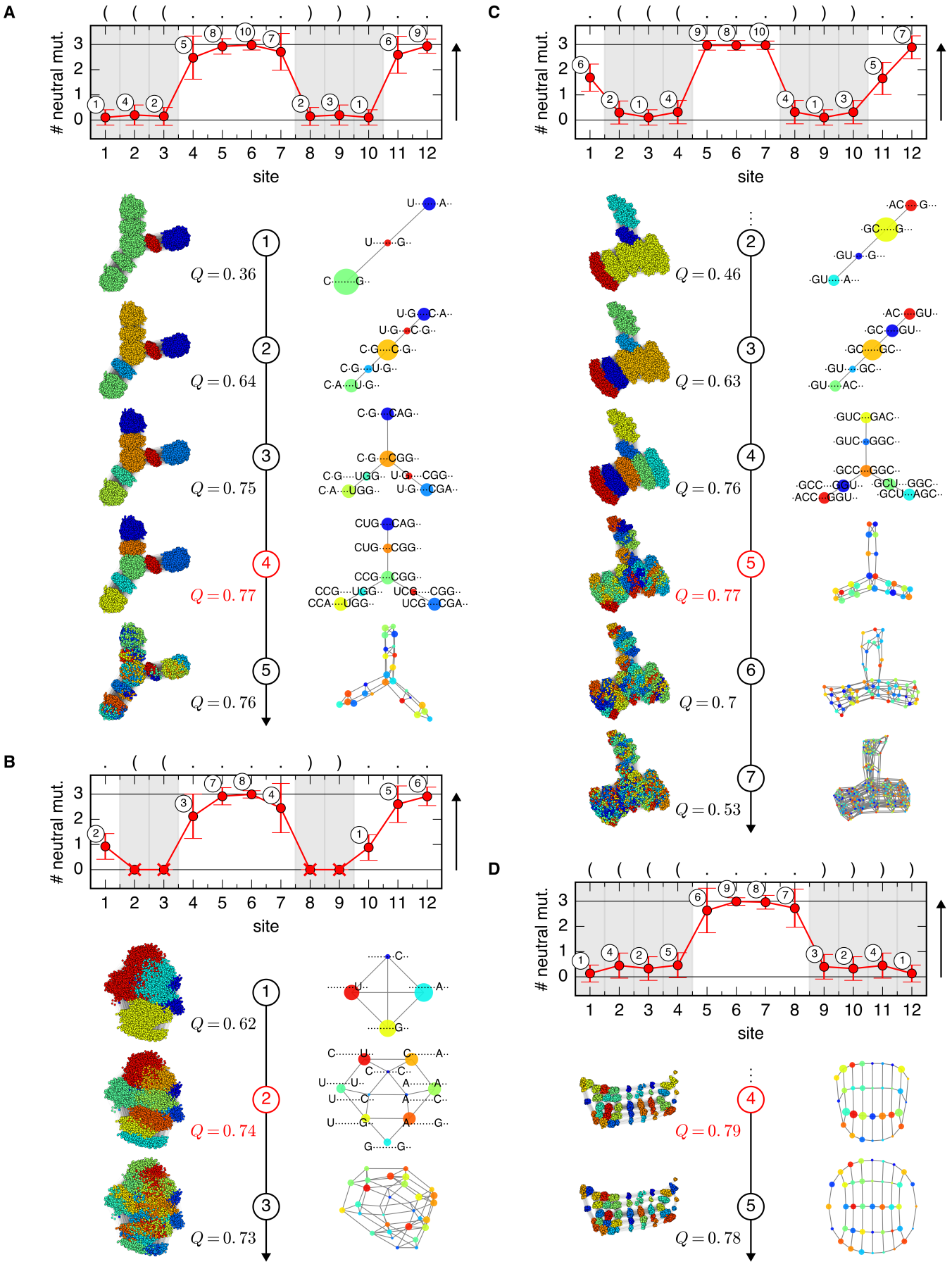
A
B
C
D
neutral mut.
site
Q = 0.36
Q = 0.64
Q = 0.75
Q = 0.77
Q = 0.76
Q = 0.62
Q = 0.74
Q = 0.73
Q = 0.46
Q = 0.63
Q = 0.76
Q = 0.77
Q = 0.7
Q = 0.53
Q = 0.79
Q = 0.78

FIG. 1. Depiction of the sequence-based communities method applied to four example NCs of the $L = 12$ RNA secondary structure GP map: (A) NC of rank 32 (size: 19488, secondary structure: '(((....)))..'), (B) 36 (size: 18468, '.((....))...'), (C) 41 (size: 16815, '.(((...)))..') and (D) 94 (size: 7341, '((((....))))'). In each of the four cases, at the top, the average number of neutral mutations per site averaged over all genotypes of the NC is displayed. The crosses indicate fully constrained sites (zero average number of neutral mutations). The shaded grey areas highlight paired sites. The numbers indicate the ordering of the sites according to their constraint (average number of neutral mutations), with the site having the smallest non-zero average number of neutral mutations receiving number 1 and so on. If sites have exactly the same constraint, they receive the same number. In each case, underneath the top figure, for a range of steps, the full NC network with coloured communities, its modularity Q and its coarse-grained network representation are shown, respectively, according to associating the communities with letter combinations at the positions of the sites with a number up to and including the respective step number. Both the full and coarse-grained networks are plotted using a force-directed graph layout algorithm. In addition, if the coarse-grained networks are not too large, the associated letter combinations are shown. The red step numbers and modularity values indicate the respective step that leads to the community structure with maximum modularity. The examples demonstrate that the community structure of a NC can be revealed by considering the sites in order of their decreasing constraint levels. Larger decreases in the constraint are associated with a change of the hierarchy layer.

i.e. whenever the position of a site from a new cluster is added, or in other words, when there is a significant change in the average number of neutral mutations. For step four, we discover seven communities like those in step four of the first example. For step six, we find 70 communities, which still largely mirror the organisation of the nodes in the force-directed layout. The maximum modularity is found for step five, which is only slightly larger than the modularity found for step four.

Finally, in the fourth example in Figure 1 (D), the NC of rank 94 is considered, which corresponds to a phenotype with four base pairs. For this example, we only show two steps as the principles are the same as before. After the fifth step, the positions of all roughly similarly constrained paired sites are added and the hierarchy layer is fully built up and matches the force-directed layout. The step afterwards leads to a change of the hierarchy layer and to a less relevant fine-grained community structure. The maximum modularity is reached for step four, but is only slightly larger than for step five.

C. Modularity comparison

Next, we benchmark our method in terms of the maximum modularity values against two common community detection algorithms in network science. For this, we consider all of the 200 largest NCs of the $L = 12$ GP map, which cover about 95.0% of the genotypes with a defined phenotype. We use this restriction since these NCs are reasonably large for a meaningful community analysis, though the exact threshold of 200 is arbitrary.

In Figure 2, for these NCs, the maximum modularity values by our method are shown versus the modularity values provided by the Louvain and spinglass algorithms, respectively. The Louvain algorithm [26] is a heuristic community detection algorithm built on an optimisation of the modularity [26]. The spinglass algorithm [27] uses a statistical mechanics approach and associates the community structure with an energy minimising spin configuration [27]. It is also one of the algorithms considered in [14] to find the 'topological communities' of the considered example NC. This algorithm is not part of the Python package *NetworkX* but the package *igraph* (also referred to as *python-igraph*), which we use in this case (version 0.7.1).

In comparison to the Louvain algorithm, our method reveals community structures with modularity values that are roughly equal or larger (for small values). Compared to the spinglass algorithm, there is a great agreement in the modularity values. This proves that our simple method is able to find meaningful community structures in RNA secondary structure NCs as well as two much more complex community detection algorithms. It should be noted that the communities found by the Louvain and spinglass algorithms do not (exactly) match those found by our method – an issue that is also discussed in [14].

III. FAST COMMUNITY SAMPLING

Below, we will introduce a method to estimate the community structure of a NC from samples of genotypes in cases when the full NC is unknown. A prerequisite for this is the ability to generate a genotype sample that spaciously covers the NC. This means that a sampling method is required that can explore the different communities in a fast way. Recently, we introduced a method to estimate the size and robustness of NCs from small samples [28]. In this context, we developed a sampling method that we refer to as 'site scanning sampling', which proved to perform significantly better than a simple random walk (RW) sampling [28]. Here, building on our understanding of the community structure, we will discuss the reasoning behind this method in more detail and will again benchmark it against RW sampling.

A. Methods

As described in [28], RW sampling starts with a random genotype on the NC and tests random one-point mutations, i.e. we randomly select a site, and randomly select a letter to which the letter at this site is mutated.

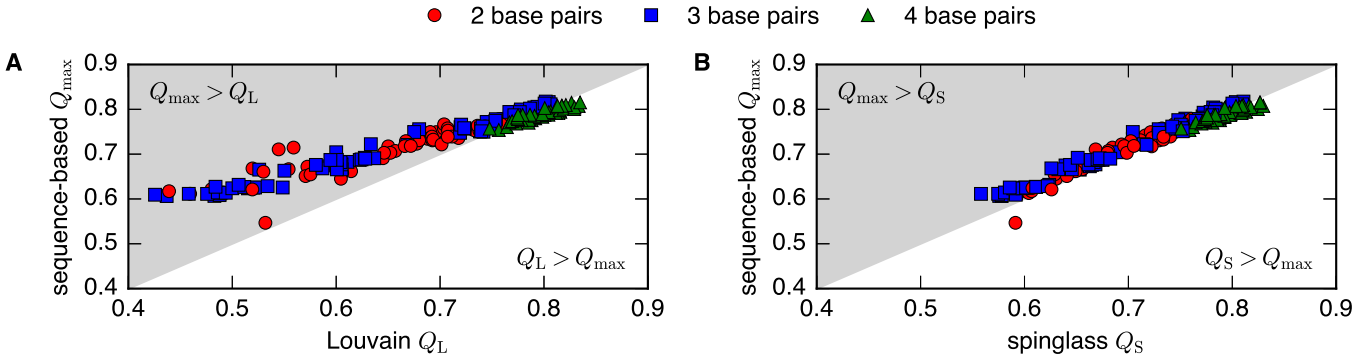

FIG. 2. Maximum modularity values $Q_{\max}$ by our sequence-based communities method versus the modularity values Q by two common community detection algorithms: (A) Louvain (Q_{L}) and (B) spinglass (Q_{S}) algorithm, for the 200 largest NCs of the $L = 12$ RNA secondary structure GP map. The coloured dots indicate the number of base pairs in the phenotypes corresponding to the NCs. Our method reveals community structures of similar or larger modularity than the Louvain algorithm, and of similar modularity to the spinglass algorithm.

If the mutation is neutral, the mutated genotype is added to the sample and serves as the new starting point, otherwise a new random one-point mutation is tested. We repeat this procedure until a sample of designated size S is reached. Since random mutations of less constrained sites are more likely to be neutral than random mutations of more constrained sites, we know from our knowledge of the NC community structure that RW sampling favours a walk within communities, rather than between them. Already by definition, RWs on networks are biased towards higher degree nodes [29]. RWs and population dynamics resembling a RW on NCs of the RNA secondary structure GP map have been studied before. Thereby, a concentration 'at highly connected parts of the network' [18] and a 'phenotypic entrapment' [19] have been observed.

In order to facilitate a walk that proceeds between communities, mutations of constrained sites need to be enforced. We achieve this with site scanning sampling, which works as described in [28]. Again, we start with a random genotype on the NC and periodically 'scan' the sequence sites from left to right. We begin with the first site and randomly mutate the letter at this site. If the mutation is neutral, the mutated genotype is added to the sample and we proceed with this new genotype and its second site. If it is not neutral, we randomly test – until we are successful – all remaining mutations of the letter at this site. If there is no success at all, the process is repeated with the initial genotype and its second site, and so on. As before, we repeat this procedure until a sample of designated size S is reached. This algorithm forces mutations of constrained sites whenever possible, but also constantly mutates the unconstrained sites in order to allow potential new mutations of constrained sites depending on the changing occupation of all sites.

For the detailed algorithms of both methods, see the Electronic Supplementary Material in [28]. For each of the 200 largest NCs of the $L = 12$ GP map, we generate 100 independent samples up to a sample size of $S = 1000$ with RW sampling and site scanning sampling, respectively.

B. Results

In Figure 3 (A), we show samples generated by the RW and site scanning approaches for each of the four example NCs from Figure 1. These results demonstrate that site scanning sampling leads to a faster exploration of the NC communities in comparison to RW sampling, which often spends longer times inside a single community.

In Figure 3 (B), we plot the average number of accessed communities as a function of the sample size, for each of the four NCs. As basis for the number of communities, respectively, we use the community structure with maximum modularity obtained by our sequence-based communities method. The results further underline that site scanning sampling outperforms RW sampling in terms of the number of accessed communities for all sample sizes for the shown four NCs. In Figure 4, we show these results in terms of the average fraction of accessed communities but additionally averaged over the 200 largest NCs. This confirms that the findings hold on average for all NCs.

In all the following sections, we employ site scanning sampling.

IV. COMMUNITY STRUCTURE ESTIMATION

In a final step, we consider the NCs of longer, naturally occurring functional non-coding RNA sequences. Such RNA sequences can be found in the functional RNA database fRNAdb [30, 31] (`http://www.ncrna.org/`, accessed on October 3, 2018). Among other information, the fRNAdb also stores a prediction of the secondary structure. However, we only take the sequence from the fRNAdb and use the secondary structure predicted by the *ViennaRNA* package since our computational analysis relies on this package. The secondary structures stored in the fRNAdb and those predicted by the *ViennaRNA* package can differ.

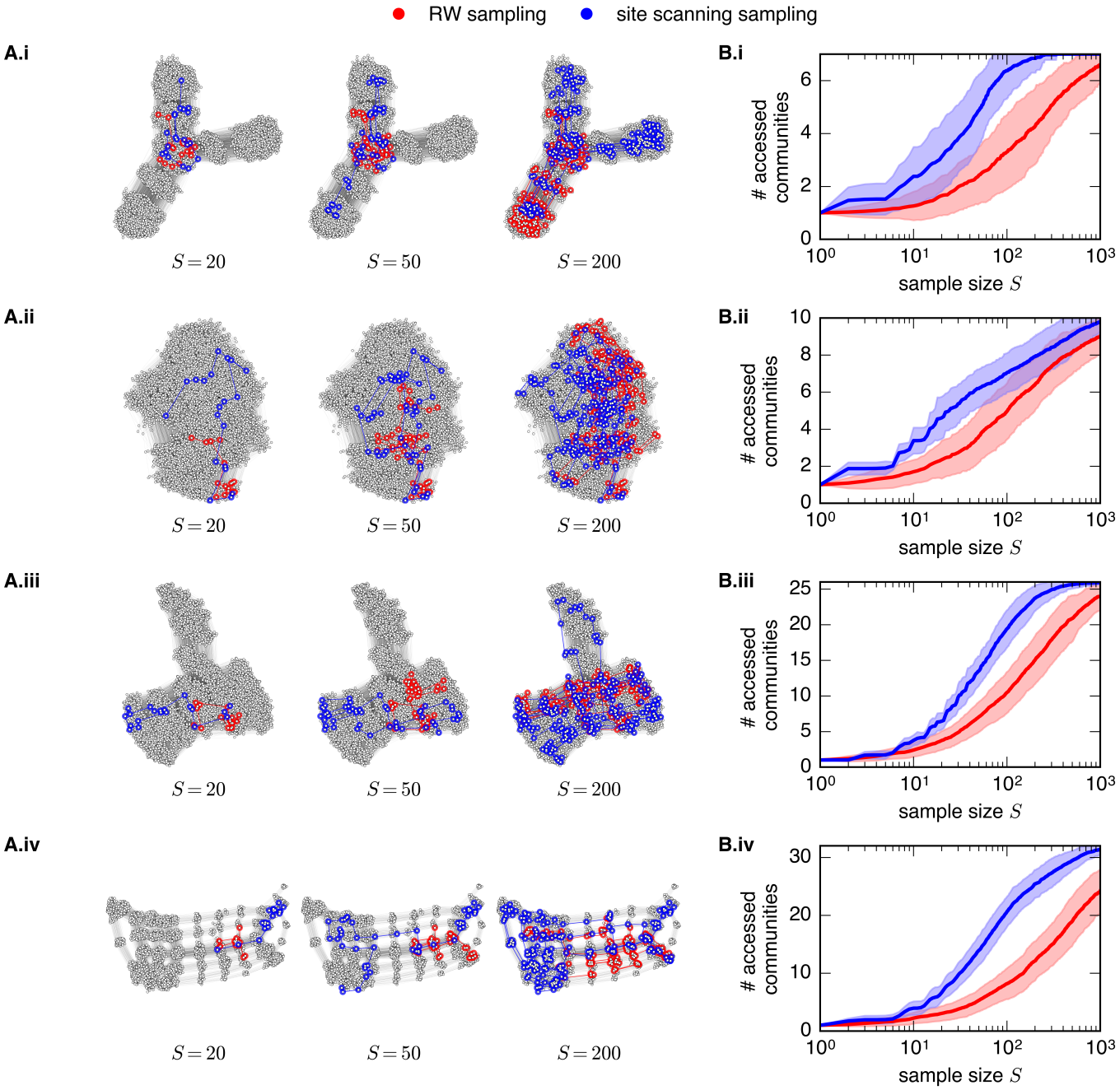

FIG. 3. (A) Examples of genotype samples of size $S = 20$, $S = 50$ and $S = 200$ generated by random walk (RW) sampling and site scanning sampling, respectively, for the four example NCs of the $L = 12$ RNA secondary structure GP map (also shown in Figure 1): (i) NC of rank 32, (ii) 36, (iii) 41 and (iv) 94. (B) Average number of accessed communities as a function of the sample size S for both sampling methods, averaged over 100 repetitions of the sampling, respectively. The shaded bands indicate the standard deviation. As basis for the number of communities, respectively, we use the community structure with maximum modularity obtained by our sequence-based communities method. In all cases, site scanning sampling leads to a faster exploration of the NC communities.

An exhaustive analysis of the NC networks of such sequences is not feasible, and their community structure has to be estimated. In this section, we introduce a way to estimate the network of communities, i.e. the coarse-grained representation, of a NC from a sample of genotypes.

A. Methods

In order to create an approximate reconstruction of the network of NC communities from a genotype sample, we require estimates of three different types of information:

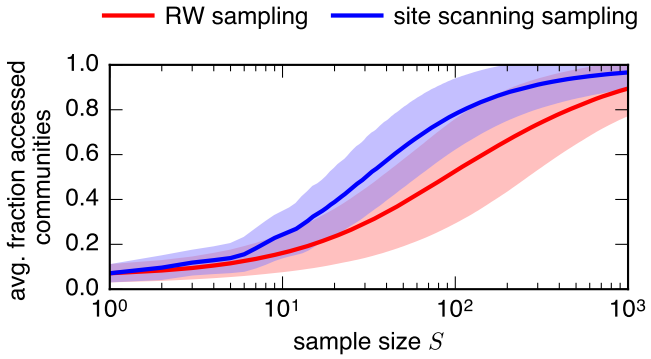

FIG. 4. Average fraction of accessed communities as a function of the sample size S for random walk (RW) and site scanning sampling, averaged over the 200 largest NCs of the $L = 12$ RNA secondary structure GP map and 100 repetitions, respectively. The shaded bands indicate the standard deviation for the averaging over the 200 largest NCs. As basis for the number of communities, respectively, we use the community structure with maximum modularity obtained by our sequence-based communities method. The results support the findings in Figure 3 (B): Site scanning sampling outperforms RW sampling in terms of a fast exploration of the NC communities.

Firstly, the list of constrained positions. Secondly, the realised letter combinations at these positions, which correspond to the communities. Thirdly, the relative frequency of the realised letter combinations.

In our original method, we found the optimal list of constrained positions by determining the average number of neutral mutations per site, and adding the positions of the constrained sites in order of their decreasing constraint to identify the community structure with maximum modularity. Here, this step-wise procedure is not feasible as it is not possible to calculate the modularity of certain community structures because the full NC is not known. In addition, longer RNA sequences exhibit a broader distribution of the average number of neutral mutations per site, which implies the possibility of multiple hierarchy layers. Thus, as a starting point, we use the paired sites for the constrained positions (as in [9]). Later, by inspecting the average number of neutral mutations per site and varying the threshold, we analyse the hierarchy layers.

We determine the average number of neutral mutations per site as follows. Starting from a fRNAdb sequence, we generate a sample of size S by using an accelerated version of the site scanning sampling, which was also introduced in [28]. Whereas site scanning sampling can be applied to any type of sequence, the accelerated version is an RNA-specific algorithm that reduces the computational costs in terms of the number of calls of the function *RNA.fold*. When a paired site is mutated, we only check the mutation for neutrality by calling *RNA.fold* if the mutated base pair is still one of the six RNA secondary structure compatible base pairs: CG, GC, AU, UA, GU and UG. Otherwise, we directly regard the mutation to be non-neutral and do not call *RNA.fold*. For the detailed algorithm of accelerated site scanning sampling, see the Electronic Supplementary Material in [28]. Now, from the sample of size S, the average number of neutral mutations per site can be calculated by averaging over the sample genotypes. However, this requires the measurement of the one-point mutational neighbourhood of each sample genotype. This can be computationally expensive, in particular for longer RNA sequences. As suggested and tested in [28], a workaround is to consider a smaller random subsample of size $S_r \leq S$ from the sample of size S and to calculate the average number of neutral mutations per site only from these S_r genotypes. As done in [28], for the measurement of the one-point mutational neighbourhoods of the S_r subsample genotypes, we also use an RNA-specific updated version by employing the principle of only checking one-point mutations of paired sites for neutrality if the mutated base pair is still compatible. For the detailed algorithms of random subsampling and the one-point mutational neighbourhood measurement, see the Electronic Supplementary Material in [28]. In a similar way, a restriction to compatible base pairs has been used in an algorithm by *Jörg et al.* [32] to estimate RNA neutral network sizes and robustness.

In order to estimate the set of letter combinations that can appear at the constrained positions, we record all letter combinations at the constrained positions across all available genotypes that we know to be part of the NC. On the one hand, these are the genotypes from the sample of size S. On the other hand, we check the neighbouring neutral genotypes that we find by measuring the one-point mutational neighbourhoods of the S_r random subsample genotypes. As we primarily consider the paired sites to be constrained, the average number of neutral mutations per site and therefore the one-point mutational neighbourhood measurement of the S_r random subsample genotypes is not necessary in the first place. However, as we will see later, including the neighbouring neutral genotypes of a random subsample will prove to be valuable in finding realised letter combinations and so coarse-grained communities.

From the relative frequency of the found letter combinations among the set of checked genotypes, we estimate the relative size of the associated coarse-grained communities.

Finally, bringing together all this information, we are able to make an estimate of the coarse-grained network representation in a similar way as for the short sequence lengths for which the full NC is known.

B. Results

The first example is an RNA sequence of length $L = 20$. It has the entry ID *FR422569* in the fRNAdb and is an RNA found in *Drosophila melanogaster*. In Figure 5, the community structure estimation results for its NC are shown. The figure also displays the predicted secondary structure in dot-bracket notation. This structure comprises six base pairs and slightly differs from the one given by the fRNAdb, which misses the outermost base pair.

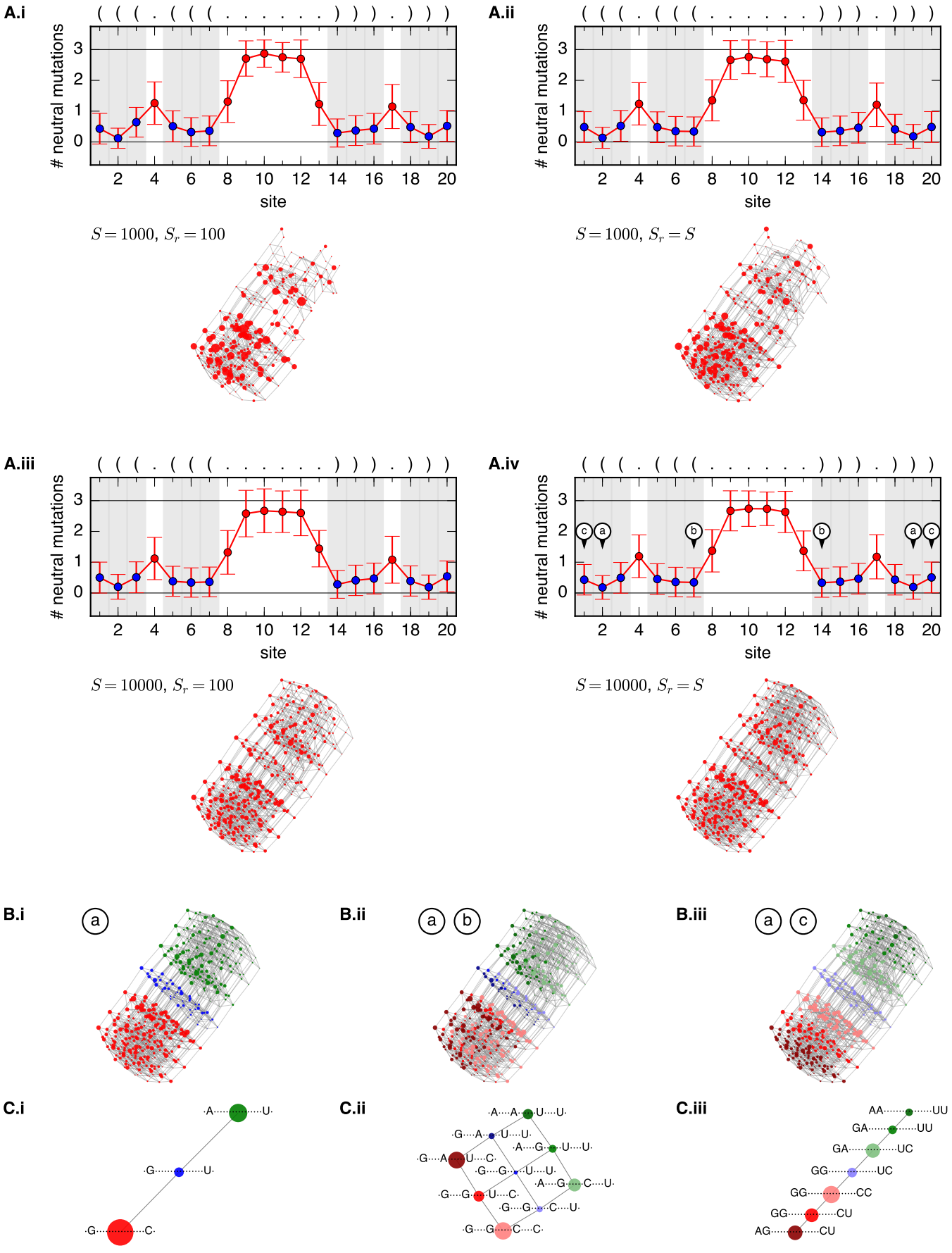
A.i
A.ii
A.iii
A.iv
(((. (((. . . . .))) .)))
neutral mutations
site
$S = 1000,\ S_r = 100$
$S = 1000,\ S_r = S$
$S = 10000,\ S_r = 100$
$S = 10000,\ S_r = S$
a
b
c
B.i
B.ii
B.iii
C.i
C.ii
C.iii

FIG. 5. (A) Community structure estimation results for the NC comprising the fRNAdb sequence with entry ID *FR422569* and length $L = 20$. For four sample and random subsample size combinations: (i) $S = 1000$, $S_r = 100$, (ii) $S = 1000$, $S_r = S$, (iii) $S = 10000$, $S_r = 100$ and (iv) $S = 10000$, $S_r = S$, the average number of neutral mutations per site averaged over the random subsample and the estimated coarse-grained network are shown. For the average number of neutral mutations per site, the shaded grey areas as well as the blue markers highlight the paired sites, i.e. the positions for which the realised letter combinations are associated with communities. For the coarse-grained network in (A.iv), a force-directed graph layout is used, the networks in (A.i), (A.ii) and (A.iii) are drawn with respect to this layout. (B) Coarse-grained network from (A.iv) with coloured communities and (C) further coarse-grained networks according to the letter combinations at positions (i) 'a', (ii) 'a' and 'b', and (iii) 'a' and 'c' marked in (A.iv), respectively. For the further coarse-grained networks, additionally, the associated letter combinations are shown. The results highlight that the coarse-grained network itself displays a community structure of which the most significant division is caused by the pair of most constrained paired sites (sites at positions 'a').

In Figure 5 (A), for four sample and random subsample size combinations, the average number of neutral mutations per site and the force-directed layout of the estimated coarse-grained network are shown. While the average number of neutral mutations per site does not differ significantly for the four combinations, the coarse-grained network builds up with increasing sample and random subsample size. The maximum shown sample size and random subsample size is $S = S_r = 10000$ (see Figure 5 (A.iv)), for which we find a coarse-grained network with a regular pattern. By using the NC size estimation formula introduced in [28], we estimate the NC size to be approximately $4.4 \cdot 10^6$. By comparing it to the number of distinct genotypes that are checked for their letter combinations at the paired sites during this sampling and estimation process, we find that about 4% of the genotypes of the NC are covered.

The coarse-grained network itself displays a community structure, which can be explained by the distribution of the average number of neutral mutations per site. In Figure 5 (A.iv), the two sites of the second outermost base pair are the most constrained. In Figure 5 (B.i), the nodes of the respective coarse-grained network are coloured based on the letter combinations at these two positions, and in Figure 5 (C.i), the related further coarse-grained network is shown. The results demonstrate that it is this pair of most constrained sites that is causing the most significant observable division of the coarse-grained network in the layout. The other paired sites have a more similar constraint. Within this group of sites, the two sites of the innermost base pair are the second most constrained ones. Figures 5 (B.ii) and (C.ii) display the coloured coarse-grained network and the further coarse-grained network, with respect to the letter combinations at the positions of the second outermost and the innermost base pair. While this additional base pair is definitely causing a division of the previously found communities into subcommunities, it is not causing the further division that is observable in the layout. As demonstrated with Figures 5 (B.iii) and (C.iii), it is the outermost base pair that is causing this observable further division. The likely reason is that the outermost and second outermost base pairs are adjacent to each other and therefore constrain the realisable letter combinations due to stability reasons, such as the avoidance of adjacent wobble base pairs.

The second example is an RNA sequence of length $L = 45$. It has the entry ID *FR039335* and is a *hammerhead ribozyme (type I)* found in *Schistosoma mansoni*. In Figure 6, the community structure estimation results for its NC are shown. The predicted secondary structure differs from the one given by the fRNAdb. It has nine base pairs in two separate stem-loops, which is not the structure associated with a *hammerhead ribozyme (type I)*. However, we will nevertheless use this example to illustrate how a coarse-grained network representation can be derived for a sequence of this length. It should also be noted that the secondary structure given by the fRNAdb would not be obtainable as a prediction from the *ViennaRNA* package. One of the stem-loops comprises no real loop, i.e. the base pair at the end of the stem spans no unpaired site, while *ViennaRNA* returns base pairs that always span at least three unpaired sites [23].

In Figure 6 (A), again, for four sample and random subsample size combinations, the average number of neutral mutations per site and the force-directed layout of the estimated coarse-grained network are shown. As before, the average number of neutral mutations per site does not change significantly between the combinations, though the coarse-grained network builds up with increasing sample and random subsample size.

The coarse-grained network again reveals a community structure that can be explained with the average number of neutral mutations per site. The sites corresponding to the left stack of three base pairs are more constrained than those of the right base pair stack. In Figures 6 (B) and (C), the coarse-grained network with coloured nodes according to the letter combinations at these positions and the related further coarse-grained network are shown for $S = S_r = 10000$ and $S = S_r = 100000$, respectively. The colouring matches the graph layouts of the coarse-grained networks underlining that the stack of more constrained base pairs is causing the observable community division. For these three base pairs, for $S = S_r = 10000$, we find ten letter combinations and so communities, while for $S = S_r = 100000$, we find four more letter combinations leading to a loop in the further coarse-grained network. This highlights that for this NC and likely for longer RNA sequences in general, more letter combinations can be found for a three base pair stack compared to the first and third example NC from Figure 1 for $L = 12$.

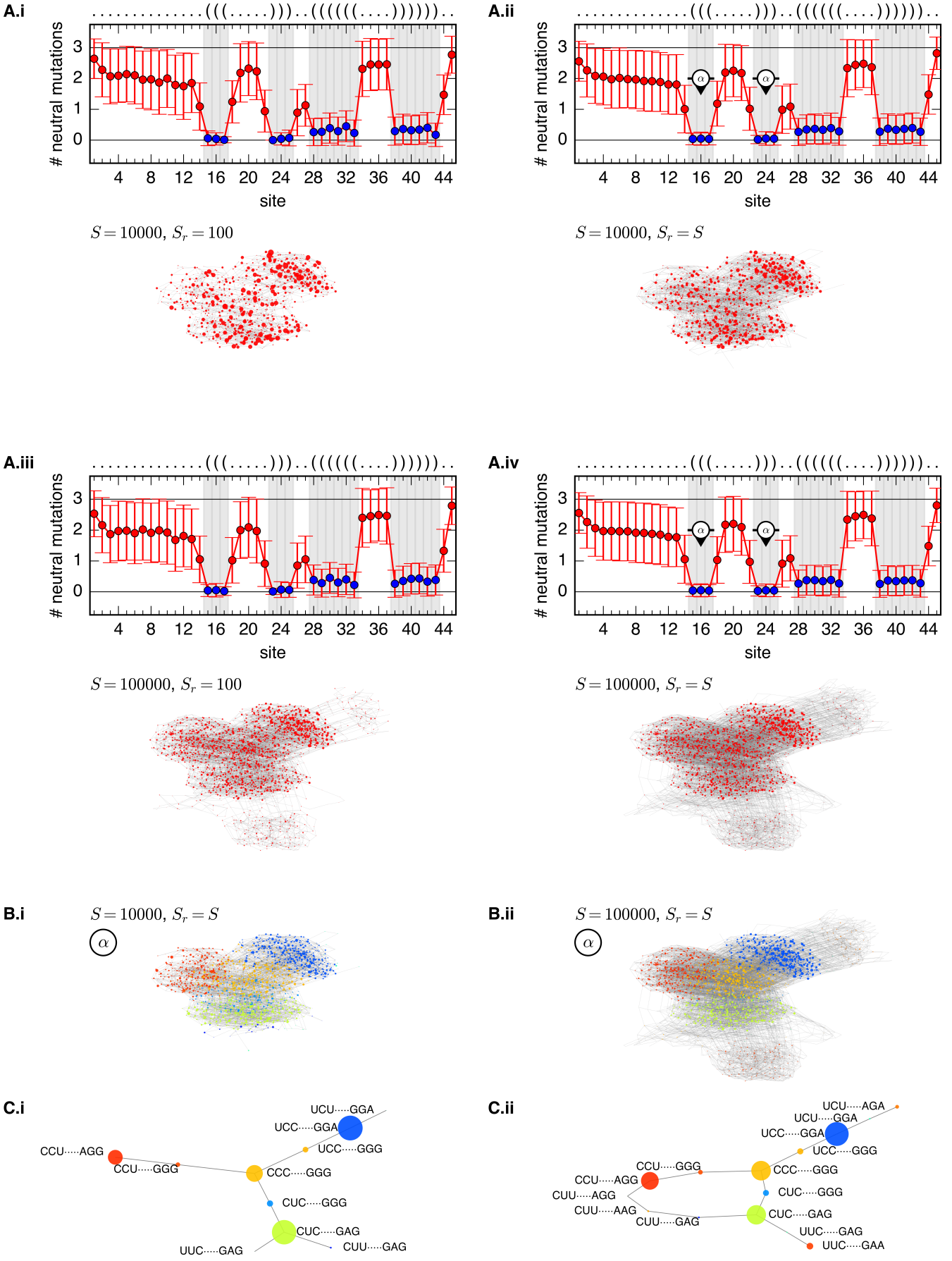
A.i
A.ii
A.iii
A.iv
B.i
B.ii
C.i
C.ii
neutral mutations
site
$S = 10000,\ S_r = 100$
$S = 10000,\ S_r = S$
$S = 100000,\ S_r = 100$
$S = 100000,\ S_r = S$
α
UCU·····GGA
UCC·····GGA
CCU·····AGG
UCC·····GGG
CCU·····GGG
CCC·····GGG
CUC·····GGG
CUC·····GAG
UUC·····GAG
CUU·····GAG
UCU·····AGA
CCU·····GGG
CUU·····AGG
CUU·····AAG
CUC·····GAG
UUC·····GAA

FIG. 6. (A) Community structure estimation results for the NC comprising the fRNAdb sequence with entry ID *FR039335* and length $L = 45$. For four sample and random subsample size combinations: (i) $S = 10000$, $S_r = 100$, (ii) $S = 10000$, $S_r = S$, (iii) $S = 100000$, $S_r = 100$ and (iv) $S = 100000$, $S_r = S$, the average number of neutral mutations per site averaged over the random subsample and the estimated coarse-grained network are shown. For the average number of neutral mutations per site, the shaded grey areas as well as the blue markers highlight the paired sites, meaning the positions for which the realised letter combinations are associated with communities. For the coarse-grained network in (A.iv), a force-directed graph layout is used, the networks in (A.i), (A.ii) and (A.iii) are drawn with respect to this layout. (B) Coarse-grained networks ((i) for $S = 10000$, $S_r = S$ from (A.ii) and (ii) for $S = 100000$, $S_r = S$ from (A.iv)) with coloured communities and (C) further coarse-grained networks according to the letter combinations at the positions marked by 'α' in (A.ii) and (A.iv). For the further coarse-grained networks, additionally, the associated letter combinations are shown. The results highlight that the most significant division of the coarse-grained network is caused by the more constrained paired sites in the left base pair stack of the secondary structure.

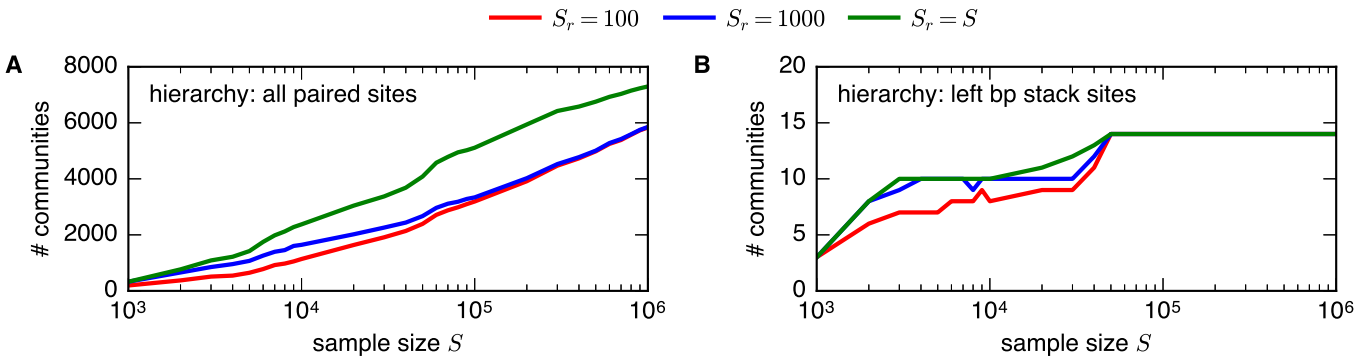

FIG. 7. Number of found coarse-grained communities of the NC comprising the fRNAdb sequence with entry ID *FR039335* and length $L = 45$ (also considered in Figure 6) as a function of the sample size S for three random subsample sizes of $S_r = 100$, $S_r = 1000$ and $S_r = S$, respectively. (A) Hierarchy layer for considering all paired sites of the respective secondary structure and (B) hierarchy layer for only considering the more constrained left base pair (bp) stack sites. While there is no saturation for the former hierarchy layer, the number of found communities saturates for the latter hierarchy layer.

For $S = S_r = 100000$, only about $2.2 \cdot 10^{-9}\%$ of the NC genotypes are covered (estimated NC size is approximately $2.4 \cdot 10^{17}$ using [28]) in the respective sampling and estimation process. It raises the question if all realised letter combinations and coarse-grained communities are found. In Figure 7, we plot the number of found communities as a function of the sample size for multiple random subsample sizes, respectively. In Figure 7 (A), the hierarchy layer for considering all paired sites is examined. In this case, no saturation is observable over the considered range of sample sizes. Furthermore, using the full sample as the random subsample ($S_r = S$) leads to significantly more discovered communities compared to using the fixed smaller random subsample sizes of $S_r = 100$ and $S_r = 1000$. In Figure 7 (B), the hierarchy layer for only considering the left base pair stack sites is examined. In this case, a saturation sets in with increasing sample size and there are less differences between the random subsample sizes. For all random subsample sizes, we find the 14 communities of this hierarchy layer for a sample size of $S = 50000$.

A potential workaround to find realised letter combinations faster could be as follows. Once an estimation with a given sample and random subsample size is completed, one could look at 'open ends' of the coarse-grained network, meaning coarse-grained nodes that are only connected to one other node (for example see Figure 6 (C.i)). By selecting a sequence from such a coarse-grained node, one could start a more specialised site scanning sampling process that still forces mutations of sites if possible but that does not allow mutations that would return to an existing coarse-grained node. This could lead to a faster discovery of unknown letter combinations and therefore of unknown coarse-grained nodes.

To summarise, the two examples show that for longer RNA sequences, the consideration of the positions of the paired sites can already lead to coarse-grained network representations that themselves show a community structure if there are differences in the site constraints of the paired sites. In this case, similar as seen for the initial sequence-based communities method, the community structure of the coarse-grained network is caused by the most constrained sites but now within all paired sites. It is probable that the longer the RNA sequence, the smaller the differences in the constraint of sites that can cause observable community divisions in the NC network.

We chose the two examples as the NCs display a particularly clear coarse-grained network representation with a further observable community division. We have also tested other fRNAdb sequences of similar lengths. For these sequences, either the paired sites were more similarly constrained (e.g. one continuous base pair stack), such that the respective coarse-grained network representation did not show a further observable community division,

caused by different site constraints along the paired sites; or the number of base pairs were larger, making the construction of the coarse-grained network computationally more expensive if all paired sites are included.

V. DISCUSSION AND CONCLUSION

Building on previous work by *Manrubia* and co-workers [9, 14], as well as by *Greenbury* [15], we introduced a method to reveal the hierarchical community structure of a NC, in the GP map of RNA secondary structure, solely from the sequence constraints and composition of the genotypes forming the NC. We identified the distribution of the average number of neutral mutations per site as the crucial starting point to identify different levels of constraint along the sequence positions. Using this knowledge, we showed that the hierarchical community layers can be revealed by proceeding through the positions in the order of their decreasing constraint and recording the realised letter combinations at these positions across the NC genotypes.

For the NCs of the $L = 12$ RNA secondary structure GP map, which are exhaustively known, we were able to find the most meaningful community structure in terms of the maximum modularity, respectively. For non-exhaustively known NCs formed by longer RNA sequences, a step-wise procedure to find the community structure with maximum modularity is not possible. Nonetheless, we outlined a way to estimate the coarse-grained community structure from a sample of genotypes. We found that observable community divisions can already be caused by differences in the constraints among the paired sites. To achieve meaningful estimates, quite large sample sizes have to be used to discover all or most of the realised letter combinations at the constrained positions. For example, for a base pair, there are three potential letter combinations within a NC. This means that the number of potential letter combinations just at the paired sites increases exponentially with the number of base pairs in the phenotype and so with sequence length. Using accelerated site scanning sampling and additionally measuring the one-point mutational neighbourhoods of random subsample genotypes have proven to improve the estimates.

Our introduced methods improve the understanding of the community structure of NCs. The community structure of NCs is likely to have an impact on evolutionary processes as has been demonstrated for other topological properties of NCs previously [18–20]. Here, we showed that a RW (which is quite similar to an evolutionary process) on a NC dominantly proceeds within rather than between individual communities as the connections between communities can be seen as 'bottlenecks'. This implies that from all alternative phenotypes surrounding a NC, which is sometimes referred to as the NC evolvability, only those surrounding the community of the starting genotype will likely to be reachable by evolution without leaving the NC, if the NC exhibits a pronounced community structure. In the future, more research should be done on the impact of the NC community structure.

We applied the framework to the NCs of the GP map of RNA secondary structure. However, the framework is not GP map specific and can almost certainly be transferred to many other GP maps. For other GP maps, the distribution of the average number of neutral mutations per site might be less bimodal than for RNA secondary structure for which paired sites are mostly constrained and unpaired sites are mostly unconstrained. Nevertheless, a similar approach to the one presented here, based on a step-wise consideration of decreasing sequence constraint, is likely to be successful in other GP maps, too.

Other GP maps this framework could be applied to include those of the HP lattice model [33–35] or the Polyomino model [12, 16], which describe different levels of protein structure. For the latter model, due to the more symmetric mapping between genotypes and phenotypes compared to RNA secondary structure, the community structures might be even more symmetric or pronounced than in RNA, as preliminary work by *Greenbury* has shown [15]. A more ambitious aim would be to apply this approach to more complex biological GP maps, such as protein secondary structure and protein tertiary structure, which are so large that sampling approaches are absolutely essential. Lastly, the approach could be applied to empirical GP maps like the one of transcription factor binding sites [17]. For this GP map, the community structure of genotype networks has been studied previously [17]. Using our sequence-based method might reveal a relationship between the communities and dual modes of binding specificity or other features of the interactions between transcription factors and binding sites [17, 36].

Author contributions

MW and SEA designed the analysis, MW carried out the analysis, MW and SEA wrote the manuscript.

Competing interests

We declare we have no competing interests.

Data accessibility

The code to generate the data shown in this article and the data itself can be accessed at: `https://github.com/mw636/NC-community.git`.

Funding

MW was supported by the EPSRC and the Gatsby Charitable Foundation. SEA was supported by the Gatsby Charitable Foundation and the Alan Turing Institute.

[1] S. Wright, Proceedings of the Sixth International Congress on Genetics **1**, 355 (1932).
[2] J. Maynard Smith, Nature **225**, 563 (1970).
[3] M. Eigen, R. Winkler-Oswatitsch, and A. Dress, Proceedings of the National Academy of Sciences **85**, 5913 (1988).
[4] S. F. Greenbury, S. Schaper, S. E. Ahnert, and A. A. Louis, PLOS Computational Biology **12**, 1 (2016).
[5] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker, Proceedings of the Royal Society of London. Series B: Biological Sciences **255**, 279 (1994).
[6] P. Schuster, Journal of Biotechnology **41**, 239 (1995), genome Research/Molecular Biotechnology, Part II.
[7] W. Fontana, BioEssays **24**, 1164 (2002).
[8] A. Wagner, Proceedings of the Royal Society B: Biological Sciences **275**, 91 (2008).
[9] J. Aguirre, J. M. Buldú, M. Stich, and S. C. Manrubia, PLOS ONE **6**, 1 (2011).
[10] E. Ferrada and A. Wagner, Biophysical Journal **102**, 1916 (2012).
[11] S. Schaper, I. G. Johnston, and A. A. Louis, Proceedings of the Royal Society B: Biological Sciences **279**, 1777 (2012).
[12] S. F. Greenbury, I. G. Johnston, A. A. Louis, and S. E. Ahnert, Journal of The Royal Society Interface **11**, 20140249 (2014).
[13] K. Dingle, S. Schaper, and A. A. Louis, Interface Focus **5**, 20150053 (2015).
[14] J. A. Capitán, J. Aguirre, and S. Manrubia, Chaos, Solitons & Fractals **72**, 99 (2015).
[15] S. F. Greenbury, *General properties of genotype-phenotype maps for biological self-assembly*, Ph.D. thesis, University of Cambridge (2014).
[16] I. G. Johnston, S. E. Ahnert, J. P. K. Doye, and A. A. Louis, Phys. Rev. E **83**, 066105 (2011).
[17] J. Aguilar-Rodríguez, L. Peel, M. Stella, A. Wagner, and J. L. Payne, Evolution **72**, 1242 (2018).
[18] E. van Nimwegen, J. P. Crutchfield, and M. Huynen, Proceedings of the National Academy of Sciences **96**, 9716 (1999).
[19] S. Manrubia and J. A. Cuesta, Journal of The Royal Society Interface **12**, 20141010 (2015).
[20] J. Aguirre, P. Catalán, J. A. Cuesta, and S. Manrubia, Open Biology **8**, 180069 (2018).
[21] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, Monatshefte für Chemie / Chemical Monthly **125**, 167 (1994).
[22] I. L. Hofacker, Nucleic Acids Research **31**, 3429 (2003).
[23] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, Algorithms for Molecular Biology **6**, 26 (2011).
[24] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[25] M. E. J. Newman, Physical Review E **69**, 066133 (2004).
[26] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Journal of Statistical Mechanics: Theory and Experiment **2008**, P10008 (2008).
[27] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).
[28] M. Weiß and S. E. Ahnert, Journal of The Royal Society Interface **17**, 20190784 (2020).
[29] N. Masuda, M. A. Porter, and R. Lambiotte, Physics Reports **716-717**, 1 (2017).
[30] T. Kin, K. Yamada, G. Terai, H. Okida, Y. Yoshinari, Y. Ono, A. Kojima, Y. Kimura, T. Komori, and K. Asai, Nucleic Acids Research **35**, D145 (2007).
[31] T. Mituyama, K. Yamada, E. Hattori, H. Okida, Y. Ono, G. Terai, A. Yoshizawa, T. Komori, and K. Asai, Nucleic Acids Research **37**, D89 (2008).
[32] T. Jörg, O. C. Martin, and A. Wagner, BMC Bioinformatics **9**, 464 (2008).
[33] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[34] D. J. Lipman, W. J. Wilbur, and J. M. Smith, Proceedings of the Royal Society of London. Series B: Biological Sciences **245**, 7 (1991).
[35] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[36] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk, Science **324**, 1720 (2009).